DataTips

Formation

# Azure Data Engineer Certification

DataTipsLearning.com

**Tarif :**
**2000 € HT**

**Durée :**
**4 Jours**
**(28h)**

**Lieu :**
**PARIS/**
**Distancielle/**
**Votre Local**

## Contact :

contact@datatipslearning.com
+33 0751386021
Réf. DTPSL-AZCDE-2022
S'inscrire

## Plus de détails :

Objectifs
Public / Prérequis
Programme détaillé
Méthode pédagogique

Azure est l'un des principaux fournisseurs de cloud au monde, fournissant de nombreux services d'hébergement et de traitement de données. Aujourd'hui, la plupart des entreprises sont natives du cloud ou migrent vers le cloud beaucoup plus rapidement. Cela a conduit à une explosion des emplois en ingénierie des données.

La certification DP-203 est un moyen infaillible de montrer aux futurs employeurs que vous avez ce qu'il faut pour devenir Azure Data Engineer.

# Objectifs

- Explorer les options de calcul et de stockage dans Azure.
- Exécuter des requêtes interactives à l'aide de pools SQL serverless.
- Effectuer une exploration et une transformation des données dans Azure Databricks.
- Explorer, transformer et charger des données dans le Data Warehouse à l'aide d'Apache Spark.
- Ingérer et charger des données dans le Data Warehouse.
- Transformer des données avec Azure Data Factory ou Azure Synapse Pipelines.
- Intégrer des données à partir de notebooks avec Azure Data Factory ou Azure Synapse Pipelines.
- Prendre en charge le traitement analytique transactionnel hybride (HTAP) avec Azure Synapse Link.
- Appliquer une sécurité de bout en bout avec Azure Synapse Analytics.
- Exécuter un traitement des flux en temps réel avec Stream Analytics.
- Créer une solution de traitement de flux avec Event Hubs et Azure Databricks.

# Public

- Data Engineer
- Data Architectes
- Data Scientist
- Data Analyst

# Prérequis

- Connaissances des bases de Azure.

# Programme détaillé

I. Les bases de Azure
    1. Présentation du portail Azure
    2. Explorer Azure accounts, subscriptions, et resource groups
    3. Présentation des services Azure
    4. Explorer Azure VMs

5. Explorer Azure Storage
6. Explorer Azure Networking (VNet)
7. Explorer Azure Compute

II. Conception d'une structure de stockage de données
    1. Conception d'un Azure data lake
    2. Sélection des bons types de fichiers pour le stockage
    3. Conception de stockage pour le requêtage efficace
    4. Conception de stockage pour le data pruning
    5. Conception de structures de dossiers pour la transformation de données
    6. Conception d'une stratégie de distribution
    7. Conception d'une solution d'archivage de données

III. Conception d'une stratégie de partition
    1. Comprendre les bases du partitionnement
    2. Conception d'une stratégie de partition pour les charges de travail analytiques
    3. Conception d'une stratégie de partition pour l'efficacité/performance
    4. Conception d'une stratégie de partition pour les fichiers
    5. Conception d'une stratégie de partition pour Azure Synapse Analytics
    6. Identification du moment où le partitionnement est nécessaire dans ADLS Gen2

IV. Conception de la couche de service
    1. Les bases de la modélisation des données et des schémas
    2. Concevoir des schémas Star et Snowflake
    3. Conception de SCD
    4. Concevoir une solution pour les données temporelles
    5. Concevoir une hiérarchie dimensionnelle
    6. Conception pour chargement incrémentiel
    7. Conception des analytical stores
    8. Conception de métastores dans Azure Synapse Analytics et Azure Databricks

V. Implémentation de structures de stockage de données physiques
    1. Utiliser Azure Synapse Analytics
    2. Implémenter la compression
    3. Implémenter le partitionnement
    4. Implémenter le partitionnement horizontal ou du sharding
    5. Implémenter les distributions
    6. Implémentation de différentes géométries de table avec les pools Azure Synapse Analytics
    7. Implémenter la redondance des données
    8. Implémenter l'archivage des données

VI. Implémentation de structures de données logiques
1. Construire une solution pour des données temporelles
2. Construire un slowly changing dimension
3. Création d'une structure de dossiers logique
4. Implémentation de structures de fichiers et de dossiers pour un requêtage et data pruning efficaces
5. Construire des tables externes

VII. Implémentation de la couche de service
1. Fournir des données dans un schéma relationnel en étoile
2. Implémentation d'une hiérarchie dimensionnelle
3. Maintenir les métadonnées

VIII. Ingestion et transformation des données
1. Transformer des données à l'aide d'Apache Spark
2. Transformer des données à l'aide de T-SQL
3. Transformer les données à l'aide d'ADF
4. Transformer les données à l'aide des pipelines Azure Synapse
5. Transformer les données à l'aide de Stream Analytics
6. Nettoyage des données
7. Normalisation et dénormalisation
8. Transformer les données en utilisant Scala
9. Splitting des données
10. Shredding JSON
11. Encoding et decoding des données
12. Configuration de la gestion des erreurs pour la transformation
13. Effectuer une Exploratory Data Analysis (EDA)

IX. Concevoir et développer une solution de traitement en batch
1. Conception d'une solution de traitement en batch
2. Développement de solutions de traitement en batch avec Data Factory, Data Lake, Spark, Azure Synapse Pipelines, PolyBase, and Azure Databricks
3. Création de pipelines de données
4. Intégration de notebooks Jupyter/Python dans un pipeline de données
5. Concevoir et mettre en œuvre des chargements de données incrémentiels
6. Concevoir et développer des slowly changing dimensions
7. Gestion des duplicats
8. Traitement missing data
9. Traitement late-arriving data
10. Upserting data
11. Regressing to a previous state
12. Présentation d'Azure Batch
13. Configuration batch size
14. Scaling resources
15. Configuration batch retention

16. Concevoir et configurer la gestion des exceptions
17. Gestion des exigences de sécurité et de conformité

X. Conception et développement d'une solution de traitement en streaming
    1. Conception d'une solution de traitement en streaming
    2. Développement d'une solution de traitement en streaming à l'aide d'ASA, d'Azure Databricks et d'Azure Event Hubs
    3. Traitement des données à l'aide de Spark Structured Streaming
    4. Monitoring des performances et des régressions fonctionnelles
    5. Traitement des données time series
    6. Concevoir et créer des windowed aggregates
    7. Concevoir et configurer la gestion des exceptions
    8. Upserting data
    9. Concevoir et créer des tests pour les pipelines de données
    10. Configuration des checkpoints/ watermarking
    11. Replaying archived stream data
    12. Transformations à l'aide de streaming analytics
    13. Gérer schema drifts
    14. Traitement sur plusieurs partitions
    15. Traitement dans une partition
    16. Scaling resources
    17. Gestion des interruptions
    18. Optimisation des pipelines à des fins analytiques ou transactionnelles

XI. Gestion des batches et des pipelines
    1. Triggering batches
    2. Gestion des chargements des loads Batch échoué
    3. Validation des loads Batch
    4. Planification des pipelines dans Data Factory/Synapse pipelines
    5. Gestion des pipelines dans Data Factory/Synapse pipelines
    6. Gestion des jobs Spark dans un pipeline
    7. Implémentation du contrôle de version pour les artefacts de pipeline

XII. Conception de la sécurité
    1. Présentation des exigences de sécurité et de confidentialité
    2. Concevoir et implémenter le chiffrement des données pour les données au repos et en transit
    3. Concevoir et implémenter une stratégie de masquage des données
    4. Concevoir et implémenter du contrôle d'accès basé sur les rôles Azure et d'une liste de contrôle d'accès de type POSIX pour Data Lake Storage Gen2
    5. Concevoir et implémenter la sécurité au niveau des lignes et des colonnes
    6. Concevoir et implémenter une politique de conservation des

données
7. Conception pour purger les données en fonction des besoins
8. Concevoir et implémenter une stratégie d'audit de données
9. Gestion des identités, des clés et des secrets sur différentes technologies de plate-forme de données
10. Implémenter des endpoints sécurisés (privés et publics)
11. Implémenter des jetons de ressources dans Azure Databricks
12. Charger un DataFrame avec des informations sensibles
13. Écriture de données chiffrées dans des tables ou des fichiers Parquet
14. Concevoir data privacy et gérer les informations sensibles

XIII. Monitoring du stockage et du traitement des données
1. Implémenter logging par Azure Monitor
2. Configuration des services de monitoring
3. Comprendre les options personnalisées de logging
4. Interprétation des métriques et des logs Azure Monitor
5. Monitoring et mise à jour des statistiques sur les données dans un système
6. Mesurer les performances des requêtes
7. Mesurer les performances du mouvement des données
8. Monitoring les performances du pipeline de données
9. Interpréter un DAG Spark
10. Monitoring les performances du cluster
11. Planification et monitoring des tests de pipeline

XIV. Optimisation et troubleshooting du stockage et du traitement des données
1. Compactage de petits fichiers
2. Réécriture des fonctions UDF
3. Gérer skews dans les données
4. Gérer data spills
5. Tuning shuffle partitions
6. Trouver le shuffling dans un pipeline
7. Optimisation des pipelines pour des workloads descriptive vs workloads analytical
8. Optimiser la gestion des ressources
9. Réglage des requêtes à l'aide d'indexeurs
10. Réglage des requêtes à l'aide du cache
11. Optimisation des pipelines à des fins analytiques ou transactionnelles
12. Troubleshooting un job Spark échoué
13. Troubleshooting un pipeline échoué

XV. Exemples de questions et solutions

# Méthode pédagogique

La formation se compose d'une partie théorique, et également une partie pratique représentant 60% de de la formation.
La partie pratique contient plusieurs exercices sous forme de notebook Databricks avec les corrections, avec aussi un projet à la fin de la formation comme simulation d'une prod.

Chaque jour, une évaluation rapide des connaissances est effectuée avant de commencer les nouvelles parties de la formation.
A la fin, une synthèse globale est délivré aux stagiaires, renforcé par un projet prod.
Finalement, une évaluation QCM est proposée.

Un support de cours sera remis à chaque stagiaire comprenant les slides, les exercices et les corrigés et un git du projet prod.
Une feuille de présence est fournie en fin de formation avec une certificat de complétion de formation pour chaque stagiaire.

Le formateur est un Data Engineer expert, qui intervient sur le sujet depuis plusieurs années en formation mais aussi en conseil.